DOCUMENT RESUME

ED 296 003 TM 011 923

AUTHOR Perlman, Carole L.; And Others
TITLE Investigating the Stability of Four Methods for Estimating Item Bias.
PUB DATE Apr 88
NOTE 21p.; Paper presented at the Annual Meeting of the National Council for Measurement in Education (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Education; Elementary School Students; *Estimation (Mathematics); Item Banks; Minimum Competency Testing; Minority Group Children; Statistical Analysis; *Statistical Bias; *Test Bias; *Test Items; Test Reliability
IDENTIFIERS Mantel Haenszel Procedure

ABSTRACT

The reliability of item bias estimates was studied for four methods: (1) the transformed delta method; (2) Shepard's modified delta method; (3) Rasch's one-parameter residual analysis; and (4) the Mantel-Haenszel procedure. Bias statistics were computed for each sample using all methods. Data were from administration of multiple-choice items from the Chicago Minimum Proficiency Skills Item Bank to 54,896 eighth graders, including Blacks, Whites, and Hispanics. The item bank consists of about 1,000 minimum competency items from which an annual form of 63 items is assembled. This study was based on the 46 items that were common to the 1984 and 1985 forms of the test. Thirty samples of each of 600 to 2,000 subjects were randomly drawn from the study population and tested over a 2-year period. Emphasis was whether bias related to socioeconomic differences existed. When individual group sizes were 666 or larger, the reliability of item bias indicators was quite adequate. For smaller groups, it was sometimes a problem. No one method consistently flagged more or fewer items across comparisons or the sample sizes. There was general agreement among methods on the proportion of times each item was indicated as being biased. Six tables provide the study data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED296003

INVESTIGATING THE STABILITY OF FOUR METHODS FOR ESTIMATING ITEM BIAS

by

Carole L. Perlman
Nikolaus Bezruczko
Linda K. Junker
Arthur J. Reynolds
William K. Rice
E. Matthew Schulz

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
CAROLE L. PERLMAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Chicago Public Schools

Department of Research and Evaluation

A paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, April, 1988.

TM011923

ACKNOWLEDGEMENT

The authors gratefully acknowledge the assistance of Prof. Nambury S. Raju of the Illinois Institute of Technology, who provided us with a PC item bias analysis program and took the time to offer suggestions and patiently answer our questions.

INTRODUCTION

The study of differential item performance or item bias is as old as standardized testing. In 1905, in an early version of their intelligence test, Binet and Simon found group differences in test scores between working class children and those from higher social classes. Cultural bias remained an issue through the first half of the twentieth century and in 1951, Eells, et al., published their classic study, "Intelligence and Cultural Differences."

In 1968, the study of bias entered the modern era when Cleary and Hilton published their study of item types unusually difficult for minority groups. Since then, research on methods of detecting bias has increased rapidly. Techniques have multiplied and computer applications have simplified the estimation of item bias. By 1981, Shepard, Camilli, and Averill identified 16 methods or variations of methods commonly used to detect item bias.

The development and refinement of methods for detecting bias continues in contemporary research. Some current contributions are Shepard's modification of the delta method, Raju's method for computing the area between two item characteristic curves (Raju, in press), and Holland and Thayer's recommendation (1986) to apply the Mantel-Haenszel procedure to the detection of group differences in item performance.

The developments and trends in assessing bias have received impetus from courts and state legislatures throughout the country. A recent example is the enactment of legislation in New York requiring the analysis of all standardized tests for item bias.

These developments in detecting item bias appear to favor formal statistical methods rather than judgmental methods such as expert raters. Empirical research by Qualls and Hoover (1981) found that differences in item difficulty showed little or no relationship to a rater's perceptions of racial favoritism. In teacher ratings of bias for 30 items from the Iowa Tests of Basic Skills, they found intraclass reliabilities of .03 for white and .11 for black raters.

Despite the clarity and fairness promoted by objective statistical methods, the stability of item bias estimates has only recently been systematically investigated. Scheuneman (1980) appears to have expressed the first concern that estimates commonly produced by statistical bias techniques are unstable. Linn, Levine, Hastings and Wardrop (1981) came to the same conclusion in their analysis of four latent trait methods and Ironson (1982) concluded that "we need more information on the reliability of the methods" (p. 152).

In the most recent study of item bias reliability, Hoover and Kolen (1984) found that six methods in common use for estimating bias "may not lead to reliable decisions about bias" (p. 180). They suggested further studies with larger sample sizes and test items that are more clearly subject to an influence of bias. The suggestions by Hoover and Kolen and concerns by other researchers are the bases for this reliability study of item bias estimates.

METHOD

The data for this study came from the administration of multiple-choice test items from the Chicago Minimum Proficiency Skills Item Bank (Citywide Report 87-1, 1987). This item bank consists of approximately 1,000 minimum competency items from which an annual form of 63 items is assembled. The dichotomously scored items are structured into a three-subscale test (language arts, computation, problem solving) and administered to 30,000 eighth graders. The test is generally characterized by an alpha reliability of approximately .90.

Items administered prior to 1986 were used in this study because a statistical method for estimating item bias was not yet in use, although bias review panels were used. This study was based on the 46 items that were common to the 1984 and 1985 forms of the test.

Hoover and Kolen suggested that reliabilities of bias indices be examined with larger sample sizes than they used in their study (200 per group). With both sexes and three racial groups evenly distributed within each sample, 30 samples each of sizes 600 and 2,000 were randomly drawn from a population of 54,986 students who were tested with 46 common items over a two-year period. Sampling without replacement was done insofar as possible, although it was necessary to do a small amount of sampling with replacement in order to obtain adequate *N*s in each cell of the larger samples.

Based upon either their wide use in the field or their promise for future use, the following four methods were compared:

- Transformed delta method
- Shepard's modified delta method
- Rasch's one-parameter residual analysis
- Mantel-Haenszel procedure

Bias statistics for each sample were computed using all four methods. Rasch item bias analyses were per performed with the programs MSCALE (Wright, et al., 1987) and LINK (Schulz, 1984) on an IBM 3033 computer. All other analyses were run on an XT-compatible PC using an untitled computer program by Raju.

Item bias indices of all methods were converted to z-scores in order to simplify comparisons. Rasch item bias indices were already z-scores, being the standardized differences between paired item calibrations. Square roots of Mantel-Haenszel chi-squares were taken and signs indicating the direction of bias were attached. Delta and modified delta indices were converted to z-scores by dividing each index by the within items standard deviation of the index across the 30 samples. Because of the relatively large number of significance tests involved in this project, an alpha of .01 was used to identify biased items.

Each set of 30 samples yielded a 46 by 30 matrix of z-scores. Error variance was estimated by the pooled variance of item bias within rows (items), across replications. Total variance was estimated by the pooled variance of bias indices within columns (samples), across items. Reliability coefficients were calculated as the ratio of true variance (total variance minus error variance) to total variance. The reliability coefficients predict the correlation between paired item-bias indices across any two samples.

RESULTS

Mean test scores, standard deviations, and Ns are given in Table 1 for males, females, and members of the three racial groups.

Insert Table 1 about here

The total N for large samples (those of size 2,000) exceeds the population size of 54,896 because some sampling with replacement was necessary in order to obtain equal numbers of blacks, whites, and Hispanics (no sampling with replacement was required to obtain the smaller samples.

The proportion of times an item was flagged as biased by each method and intergroup comparison is given in Table 2 for the 30 small samples (total N = 600) and in Table 3 for the 30 large

---

Insert Tables 2 and 3 about here

---

samples (total N = 2,000). As would be expected given the increased power of the z-test at the larger sample size, more items were flagged as being biased using samples of size 2,000 rather than 600. No one method consistently flagged more or fewer items across the various comparisons or across the two different sample sizes. With few exceptions, there was agreement among methods on the proportion of times each item was indicated as being biased.

Table 4 displays the summary statistics for each item bias method by sample group. Two results are apparent. First, although

---

Insert Table 4 about here

---

mean item values are consistent in large and small samples across bias methods, standard deviation and range values differ markedly. Small sample standard deviations for each method are approximately one-half of those in the larger sample. This indicates the relative lack of sensitivity in detecting item bias in small samples (individual group sizes of 300 or less). Second, with the possible exception of the male/female contrast, the methods appear to be similar in their sensitivity of detecting bias as standard deviation and range values are fairly similar in most intergroup comparisons. With the largest individual group sizes (666 or more per group when N=2000), Mantel-Haenszel and Rasch methods were particularly sensitive in detecting bias.

Table 5 contains reliability estimates for the item bias statistics. The reliabilities for the different methods were

---

Insert Table 5 about here

---

comparable at each group size, although the modified delta reliabilities were generally slightly lower than the others. Only when the number of cases in each comparison group reached 1,000 did the reliabilities approach .80.

Table 6 indicates the high similarity of item bias methods across contrast groups aggregated over 30 samples. The diagonal

---

Insert Table 6 about here

---

elements of the matrices give an estimate of the correlation of mean item bias indices within methods. A mean item bias index is its average over the 30 samples. Thus, if the study were repeated with another 30 samples, the correlations of the mean item bias indices would be expected to be the values given on the diagonals of the matrices in Table 6. This is also the projected correlation when sample sizes are 30,000, 20,000, 9,000, or 6,000 (30 times the sample size for individual cells. The correlations in Table 6 range from .66 (small sample, group1/group3, modified delta/Rasch) to 1.00 (small and large samples, male/female, Mantel-Haenszel/Rasch). The modified delta method consistently has the lowest correlations with other methods.

DISCUSSION

The reliabilities and inter-method correlations presented in this study are considerably higher than those obtained by Hoover and Kolen, although reliabilities at smaller sample sizes (200-300 per comparison group) are hardly impressive. Since the findings are necessarily test-specific and sample dependent, it is not surprising that the two studies yielded somewhat different results. Although both studies indicate that reliability can definitely be problem with smaller samples, the current study demonstrates that

when individual comparison group sizes are 666 or more, reliability of item bias indicators can be quite adequate. However, sample sizes that large are not always available.

Test developers who choose to use statistical methods for assessing item bias may wish to adopt one or more of the following strategies:

- Run more that one type of statistical bias analysis and compare the results obtained for each item. All of these analyses may be performed on a PC using relatively inexpensive programs.
- Compare results from a statistical analysis with results from a bias screening panel. This is hardly a new idea, but it still makes good sense.
- If there are adequately many items and the cost of having biased items is high, use a high alpha level, especially if the sample size is small.
- If the items are given over time to more than one group of people, look for stability of bias indices over successive administrations of the items.

Obviously, there are still questions that are left unanswered. For example, this study did not attempt to explain the source of bias for flagged items. Examining the characteristics of items flagged consistently over methods (e.g., by factor analysis) might help to identify underlying biasing characteristics. Also, given a test whose items have known bias, how will the bias indices behave under various conditions? Or what happens when the analyses are

done on randomly equivalent groups rather than on a target and comparison group? Would the error variance of the bias indices be the same for biased items as for unbiased items? These are all questions whose answers might shed some light on the usefulness of statistical indicators of item bias.

REFERENCES

Citywide Report 87-1 (1987). Minimum Proficiency Skills Test: 1987 item pilot report. Bureau of Citywide Testing: Chicago Public Schools.

Cleary, T. A. and Hilton, T. L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Eells, K. A., Davis, R. J., Havighurst, R. J., Herrick, V. E., and Tyler, R. W. (1951). Intelligence and Cultural Differences. Chicago: University of Chicago Press.

Holland, P. W. and Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the American Educational Research Association Annual Meeting, San Francisco, California, April 1986. (Revised April 22, 1986).

Hoover, H. D. and Kolen, M. J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk, (Ed.) Handbook of methods for detecting test bias (pp. 117-160). Baltimore: The Johns Hopkins University Press.

Linn, R. L., Levine, M. V., Hastings, C. N., and Wardrop, J. L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Qualls, A. and Hoover, H. D. (1981, April). Black and white teacher ratings of elementary achievement test items for potential race favoritism. Paper presented at the Annual Convention of the American Educational Research Association, Los Angeles.

Raju, N. S. (1987). The area between two item characteristic curves. Psychometrika, in press.

Schulz, E.M. (1984) LINK [A FORTRAN program for comparing paired Rasch estimates and linking tests through common items or persons]. Chicago: MESA Press.

Scheuneman, J. (1980, April). Consistency across administrations of certain indices of bias in test items. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Shepard, L. Camilli, G., and Averill, M. (1980, April). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Wright, B.D., Schulz, E. M., Congdon, R. T., and Rossner, M. (1987). MSCALE [A FORTRAN program for comparing paired Rasch estimates and linking tests through common items or persons]. Chicago: MESA Press.

Table 1

Test Score Means and Standard Deviations of Sample Groups By Gender and Race

	Sample Size					
	Large Samples Sample N = 2,000			Small Samples Sample N = 600		
Group	N	$\bar{X}$	SD	N	$\bar{X}$	SD
Gender						
Male	30,000	32.10	8.69	9,000	32.13	8.66
Female	30,000	32.13	8.29	9,000	32.11	8.29
Race						
Group 1	19,980	35.68	7.40	6,000	35.67	7.44
Group 2	20,040	30.06	8.22	6,000	30.02	8.21
Group 3	19,980	30.61	8.65	6,000	30.66	8.59
Total	60,000	32.12	8.49	18,000	32.12	8.48

Note. Test included 46 items. Samples were drawn from over 54,896 students who took the test in 1984 and 1985. Most sampling was done without replacement (see text).

Table 2

Percent of Small Samples Flagged for Bias
By Method and Item

	Contrast											
	Male/Female				Group 1/Group 2				Group 1/Group 3			
Item	D	MD	MH	R	D	MD	MH	R	D	MD	MH	R
1	03	03	03	03	00	00	00	00	00	00	00	00
2	07	07	03	07	03	03	03	03	00	00	00	00
3	00	03	03	03	00	03	00	00	03	03	00	00
4	07	37	03	03	00	10	23	00	00	03	13	00
5	07	07	00	00	20	10	00	10	10	07	00	27
6	07	07	00	00	10	10	00	10	07	03	00	07
7	03	03	03	03	00	00	00	00	10	13	03	03
8	10	13	00	00	00	03	00	00	00	00	00	00
9	00	00	00	00	00	00	00	00	00	03	03	03
10	07	03	07	10	03	03	03	03	03	03	07	07
11	00	07	00	03	00	03	00	00	00	03	00	00
12	07	03	17	20	00	00	03	00	30	00	03	23
13	30	27	07	07	03	03	00	10	10	03	17	50
14	07	03	00	00	10	03	03	07	10	03	00	03
15	03	03	23	30	03	03	00	03	23	17	10	10
16	03	03	00	00	00	00	00	00	03	03	00	00
17	00	00	03	07	00	00	00	00	10	00	17	13
18	27	00	17	17	00	00	00	00	03	03	00	00
19	00	00	00	00	00	00	00	03	10	13	07	10
20	07	07	07	07	00	00	00	00	03	03	07	10
21	00	00	00	00	33	27	00	23	07	03	03	20
22	17	13	13	07	03	03	00	00	03	00	00	00
23	07	10	00	00	00	00	00	00	03	00	00	00
24	03	07	07	00	00	00	00	00	00	03	00	00
25	00	00	03	03	33	00	10	00	50	00	63	37
26	03	03	03	03	00	00	00	00	03	03	00	03
27	23	27	30	23	03	10	00	03	07	13	00	07
28	03	03	00	00	00	00	00	00	00	03	00	00
29	00	00	00	00	30	17	13	27	03	03	03	07
30	00	03	00	00	07	03	03	03	03	00	00	00
31	00	00	00	00	07	03	03	03	03	00	00	00
32	07	07	17	13	00	00	00	00	00	00	00	00
33	03	03	03	03	07	03	03	00	57	00	47	43
34	00	00	00	00	00	00	00	00	00	00	00	00
35	13	10	40	60	00	03	03	00	00	03	00	13
36	10	13	03	07	07	03	03	07	00	00	00	00
37	17	07	20	07	00	03	03	00	00	00	00	00
38	13	03	13	13	07	10	03	07	07	03	00	07
39	07	07	10	10	00	00	00	00	00	03	00	00
40	13	13	43	57	00	00	30	00	10	00	27	07
41	00	00	00	00	10	00	00	03	00	00	00	00
42	27	27	43	37	00	00	00	00	03	10	00	00
43	00	03	00	03	10	03	10	07	03	03	00	00
44	07	07	00	03	00	00	00	00	00	00	00	03
45	03	03	07	03	00	00	00	00	07	03	00	00
46	00	00	00	00	00	00	07	07	07	03	03	03
Mean	7	7	8	8	5	3	3	3	7	3	5	7

[1]Leading decimal for proportion was removed.

Table 3

Percent of Large Samples Flagged for Bias Across Methods by Item

Item	Contrast											
	Male/Female				Group 1/Group 2				Group 1/Group 3			
	D	MD	MH	R	D	MD	MH	R	D	MD	MH	R
1	23	10	10	33	00	00	00	00	00	00	03	00
2	03	03	03	03	00	00	00	00	03	03	03	03
3	03	03	07	20	07	03	07	03	00	00	03	03
4	40	83	07	03	07	10	83	07	10	80	67	00
5	10	17	13	13	47	30	10	50	67	23	10	67
6	00	03	00	00	20	17	07	40	17	10	07	50
7	00	00	30	30	00	00	03	03	20	30	07	10
8	43	40	07	07	03	03	03	03	00	00	07	00
9	00	00	03	07	03	03	03	03	03	10	03	00
10	80	80	53	63	13	47	40	17	07	13	37	13
11	03	03	00	00	00	10	17	00	00	07	03	00
12	17	10	77	73	00	00	13	00	73	47	57	93
13	90	97	20	13	70	40	00	23	40	20	40	93
14	37	27	03	07	13	03	03	17	13	07	03	10
15	80	87	87	93	03	03	00	03	60	47	50	33
16	00	00	00	00	13	13	13	17	00	00	00	03
17	00	00	00	00	03	03	00	13	47	57	63	50
18	90	00	50	57	03	03	07	07	10	03	00	03
19	00	00	00	03	00	03	00	03	73	70	60	77
20	07	17	40	63	00	00	03	00	20	27	07	30
21	03	07	07	03	97	90	07	77	73	40	07	60
22	47	20	43	20	00	03	00	00	17	07	07	07
23	07	17	07	07	00	00	03	00	17	07	03	20
24	40	13	47	30	00	03	03	00	03	13	01	00
25	03	03	07	00	20	40	30	10	47	00	97	70
26	03	13	10	03	07	00	00	03	20	10	00	17
27	73	60	93	93	10	13	17	10	33	37	13	17
28	03	03	00	00	00	00	00	00	07	07	10	10
29	00	00	13	13	73	60	50	67	10	10	20	13
30	00	00	00	00	40	10	10	17	10	00	00	00
31	00	00	00	00	13	10	10	07	03	00	03	03
32	13	20	83	73	00	00	03	00	47	33	00	00
33	07	07	17	10	10	10	07	07	90	00	00	100
34	03	03	00	00	03	03	00	00	00	00	00	00
35	47	30	100	100	00	03	30	00	13	00	00	50
36	33	47	17	17	20	07	13	23	00	00	00	00
37	13	10	60	53	03	03	07	00	03	00	00	07
38	77	73	53	67	13	10	03	17	10	07	00	10
39	30	23	80	73	00	00	00	00	03	07	00	00
40	27	17	97	100	03	17	57	07	87	93	77	07
41	03	00	03	00	43	10	10	47	03	00	00	07
42	73	73	97	97	10	07	17	10	57	60	03	00
43	03	03	07	10	40	20	33	20	03	00	03	10
44	10	13	23	30	00	00	03	00	03	03	03	07
45	47	30	30	23	07	13	07	03	03	03	00	03
46	00	03	17	27	07	03	00	03	17	10	00	00
Mean	24	21	29	29	14	11	12	12	23	17	15	21

[1]Leading decimal for proportion was removed.

Table 4

Mean z-score Item Bias Statistics
By Method and Group Contrasts

	Contrast								
	Male/Female			Group 1/Group 2			Group 1/Group 3		
Method	$\bar{X}$	SD	RANGE	$\bar{X}$	SD	RANGE	$\bar{X}$	SD	RANGE
	Large Sample (N = 2000)								
Delta	.01	1.92	8.00	.03	1.48	6.89	.05	1.79	6.95
Modified Delta	.05	1.81	7.85	.05	1.35	6.08	.17	1.62	7.55
Mantel-Haenszel	.16	2.16	9.20	-.19	1.30	5.87	-.26	1.66	7.24
Rasch	-.11	2.20	9.34	.02	1.30	5.35	.12	1.79	8.27
	Small Sample (N = 600)								
Delta	.03	1.07	4.14	.00	.82	3.72	-.03	.99	4.13
Modified Delta	.09	.98	3.44	.09	.71	2.97	.25	.64	3.21
Mantel-Haenszel	.09	1.12	4.85	-.09	.72	3.11	-.15	.89	3.98
Rasch	-.06	1.18	5.01	-.02	.75	3.11	.08	.99	4.66

TABLE 5

Reliability of Item Bias Statistics

Model (1)	Individual Cell Size (2)	Groups (3)	Pooled Within Row Variance (4)	Pooled Within Column Variance (5)	Across Sample Reliability (6)
Delta	1000	M/F	1.00	4.64	.79
Mod-Del	1000	M/F	1.00	4.24	.76
M-H	1000	M/F	.99	5.62	.82
Rasch	1000	M/F	.97	5.76	.83
Delta	300	M/F	1.00	2.10	.53
Mod-Del	300	M/F	1.00	1.93	.48
M-H	300	M/F	.90	2.13	.58
Rasch	300	M/F	.94	2.31	.59
Delta	666	1/2	1.00	3.16	.68
Mod-Del	666	1/2	1.00	2.79	.64
M-H	666	1/2	.91	2.57	.65
Rasch	666	1/2	.94	2.59	.64
Delta	200	1/2	1.00	1.64	.39
Mod-Del	200	1/2	1.00	1.47	.32
M-H	200	1/2	.80	1.30	.38
Rasch	200	1/2	.89	1.44	.39
Delta	666	1/3	1.00	4.16	.76
Mod-Del	666	1/3	1.00	3.59	.72
M-H	666	1/3	.90	3.63	.75
Rasch	666	1/3	.96	4.12	.77
Delta	200	1/3	1.00	1.95	.49
Mod-Del	200	1/3	1.00	1.38	.27
M-H	200	1/3	.83	1.59	.48
Rasch	200	1/3	.97	1.95	.50

Note: Reliabilities are derived from 30 estimates of item bias on 46 items.

Table 6

Correlations Between Item Bias Methods
Across Samples Sizes and Group Contrasts

Cell Sample Size		Male/Female				Group 1/Group 2				Group 1/Group 3			
		D	MD	MH	R	D	MD	MH	R	D	MD	MH	R
	D					.95	.90	.80	.99	.97	.77	.91	.90
	MD						.93	.90	.90		.92	.73	.66
200	MH							.95	.82			.97	.90
	R								.95				.97
	D	.97	.90	.88	.88								
	MD		.97	.85	.85								
300	MH			.98	1.00								
	R				.98								
	D					.98	.92	.75	.98	.99	.87	.88	.89
	MD						.98	.89	.92		.99	.80	.69
666	MH							.98	.81			.99	.89
	R								.98				.99
	D	.99	.94	.87	.86								
	MD		.99	.83	.82								
1000	MH			.99	1.00								
	R				.99								

Note. Coefficients were obtained by computing the mean bias value for each item over 30 samples. The correlation between methods was then computed with items as the unit of analysis.
Diagonal elements are estimated intramethod reliabilities based on total variance from Table 4 error variance from Table 5.

Note: The following abbreviations are used in some of the tables in this section:

D = Delta

MD = Modified Delta

MH = Mantel Haenszel

R = Rasch